The Moral Responsibility Gap and the Increasing Autonomy of Systems

Zoë Porter[1], Ibrahim Habli[1], Helen Monkhouse[2], and John Bragg[3]

[1] University of York
zoe.porter@york.ac.uk
ibrahim.habli@york.ac.uk
[2] HORIBA MIRA
helen.monkhouse@horiba-mira.com
[3] MBDA UK Ltd.
john.bragg@mbda.co.uk

Abstract. The ethical and social implications of autonomous systems are forcing safety engineers and ethicists alike to confront new questions. This paper focuses on just one of these questions - moral responsibility - bringing together inter-disciplinary insights to an issue of growing public and regulatory concern. The central thesis is that, on a conception of moral responsibility that presupposes control, the increasing autonomy of systems *prima facie* diminishes the extent to which engineers and users can be considered morally responsible for system behaviour. This challenge to our normal attributions of moral responsibility as a result of autonomy has come to be known as the 'responsibility gap'. We provide a characterisation of the moral responsibility gap, which we argue has two dimensions: causal and epistemic. At the end of the paper we highlight considerations for future work.

Keywords: Moral Responsibility, Ethics, Autonomous Systems, Safety

1 Introduction

Given the public and regulatory concern with autonomous systems, such as autonomous vehicles, this paper is motivated by a need to locate where there is currently diminished control over, and uncertainty about, the behaviour of such systems. These considerations should help on two counts. First, to inform discussions about how far designers and engineers are - and should be - morally responsible for system behaviour. Second, to contribute to discussions about how to evaluate confidence in autonomous systems.

This paper starts with a brief, philosophical exposition of moral responsibility, elucidating the thesis that control is a necessary condition of moral responsibility (Part 2). The notion of the 'moral responsibility gap' is then introduced, with the argument that this has two dimensions: loss of causal control and loss of epistemic control (Part 3). The paper then examines the relevant differences between non-autonomous and autonomous systems with respect to the two dimensions of the responsibility gap (Parts 4 and 5). Finally, it highlights the

salient issues that the authors believe should constitute considerations for future work (Part 6).

2 Moral Responsibility

There are many senses in which the word 'responsible' is used across disciplines and in everyday discourse, but the etymology of the word indicates a common thread [1]. 'Responsible' comes from the Latin *respondeo*, and means 'to be answerable'. 'Moral responsibility' is primarily concerned with questions about when we are answerable for actions or events in such a way that we might be praised or blamed for them. This, in turn, depends on the particular relationship that obtains *"between people and the actions they perform, or between people and the consequences of their actions."* [2].

Philosophical theories of moral responsibility date back to Aristotle, who held that voluntary actions - in which the cause of the action is the agent himself or herself, and which he or she undertakes knowingly - were the only ones for which a person could be praised or blamed [3]. This has remained a deeply influential account of moral responsibility, and underpins many modern conceptions, though a radical development occurred in the 1960s with the work of Peter Strawson, who located moral responsibility not in objective conditions, such as whether the agent acted voluntarily, but in the wide variety of attitudes expressed within interpersonal relationships, according to which we praise, blame, feel gratitude, or resentment towards agents in virtue of how far we perceive them to be acting in accordance with our expectations of a reasonable degree of good will [4]. Recent accounts also differentiate between moral responsibility as accountability and moral responsibility as attributability [5].

For the purposes of this paper, we follow the supposition that it is only appropriate to hold a person morally responsible for an action or event over which they have some control, whereby they are not acting from either compulsion or ignorance [6]. It is important to note that ignorance is not always an exculpation. If one does not take sufficient care to discover whether one's actions will lead to harm, then attitudes of praise and blame are still appropriate. Negligence can be blameworthy [7].

Moral responsibility works in two directions. There is prospective responsibility, which a duty or an obligation to maintain or bring about a certain state of affairs, such as safety. And there is retrospective moral responsibility, which is accountability or liability for something that has already happened, such as an injury. The philosophical literature is dominated by a preoccupation with the latter kind of moral responsibility, often because of concerns about blaming people unfairly [2]. While our discussion will be similarly focused, we will also consider prospective responsibility, particularly as it bears on those trying to assure the future safety of autonomous systems.

We restrict the scope of our discussion in two ways. First, we limit our analysis to the human-side of the moral responsibility gap. There are interesting philosophical questions about the extent to which the computational systems

themselves might be morally responsible, but we do not consider such questions here. Second, this paper makes no claims about the legal implications of the moral responsibility gap. Moral responsibility and legal responsibility are not the same. Nonetheless, it is worth noting that moral responsibility has a greater overlap with criminal liability than with civil liability. In criminal cases, lack of moral fault tends to protect the defendant from liability, and certain central assumptions about moral responsibility are reflected in criminal law [8]. In civil cases, however, there can be liability without fault [2].

3 Responsibility Gaps

Though any delegated action - whether to an individual human delegee, to an institution, or to a machine - incurs some kind of a moral responsibility gap, since the agent to whom an action is delegated may act against the wishes of, or contrary to the expectations of, the delegator, we argue that when action is delegated to an autonomous system this gap is substantially widened.

The term 'responsibility gap' with respect to autonomous systems was first introduced by Andreas Matthias in a seminal paper [9], in which he argued that there is *"... an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to assume the responsibility for them. These cases constitute what we will call the responsibility problem."* (p. 117)

Extending and elucidating Matthias' treatment, we posit that the moral responsibility gap has two dimensions: the causal and the epistemic. The causal dimension of the gap can be thought of as diminishing control over the 'what' of system behaviour. Part of the problem in determining how to attribute moral responsibility to engineers of autonomous systems is the difficulty in tracing the nature and extent of the causal influence that they have over those systems' final capabilities. The epistemic dimension of the gap can be thought of as diminishing control over the 'how' of system behaviour; with autonomous systems, precisely how the system reaches a decision is increasingly something the human delegator cannot explain or understand. As such, another part of the problem in attributing moral responsibility to engineers and users of autonomous systems is the difficulty in determining the extent to which they might reasonably be expected to know, or at least seek to know, how the systems behave.

To clarify this difference between the causal (control over the 'what') and the epistemic (control over the 'how'), take an autonomous car. Developing overtaking capabilities falls under the causal dimension. Understanding of how the autonomous system will deploy these features falls under the epistemic dimension. A common problem in engineering is that it is hard to draw a line between the 'what' and the 'how.' This distinction is further complicated with autonomy.

It is important to note that, as with all modern technology, there is already a causal constraint on the moral responsibility of human delegators on two counts. First, because of the 'problem of many hands,' whereby multiple actors are

involved in designing the capabilities of such systems [10] [11]. Second, because of the temporal and physical remoteness of the systems actions from the original action of design, certification, and manufacture [10]. Autonomy adds a new layer of complexity for attributions of moral responsibility on engineers and users.

4 Responsibility Gaps and Non-Autonomous Systems

Traditional safety engineering mainly relies on the largely controlled nature of systems. A boundary definition is typically used to describe the system's key architectural elements, its functions, and its interfaces to other systems. Referred to by many safety standards as the Target of Evaluation (TOE), this becomes the foundation on which all subsequent safety activities are performed. The first such is the Hazard Analysis and Risk Assessment (HARA) process, which seeks to identify the potentially hazardous behaviour of the TOE, and also to classify the resulting risk. Safety functions can then be defined to mitigate the hazard risk, with the integrity level of the safety functions commensurate with that risk. Finally, testing the system's implemented behaviour against the safety requirements builds confidence that the system achieves the acceptable level of safety, as determined by the relevant safety standards or authorities.

For domains such as aviation and automotive, where many vehicles of the same type will be built and operated, a type-approval or certification process is commonly used. Like the safety standards discussed above, this approach primarily relies on determinism and predictability. This paradigm assumes that if one can satisfy the appropriate regulatory or certification body about the performance and safety of the first system of type built, then (providing an effective manufacturing process exists) the 100^{th} or 1000^{th} system manufactured will behave in the same way. In this way, it is possible to assure the safety of the fleet having only scrutinised a signal system of type.

Currently, therefore, the engineer retains a substantial level of control over both the 'what' (causal control) and the 'how' (epistemic control) of the system. Engineers and designers determine what the system's capabilities are and what it can be used for. The causal condition for moral responsibility is also met by safety engineers because, counter-factually, if a signal system type is not evaluated as acceptably safe, the fleet is not built. However, given that the system operates at a temporal and physical distance from the original actors, we might argue that engineers no longer have moral responsibility for the system's actions if there is a later intervention in the system, either by users or other parties, that they could not reasonably have foreseen and mitigated.

The epistemic dimension of moral responsibility - control over the 'how' of system behaviour - is also robustly met. Engineers can understand how a system works and why it takes the actions that it takes. There is an established framework for deliberation about the behaviour of the system and outcomes are largely predictable. Though the engineer is not proceeding from a position of absolute certainty about the behaviour of the system, there is a very high degree

of confidence that the system will perform to type, based on the behaviour of the system when tested against safety requirements.

We might reasonably ask questions about the adequacy of the test coverage measures themselves, and how far these are within the control of the engineers who are mandated to use them (by national or international standards). However, they exist and some are mandated by the relevant authorities.

There are also some marginal epistemic constraints, such as the possibility that an engineer might not fully understand something about the boundary of definition, or fail to imagine and identify potentially hazardous behaviour of the system in the HARA process. In both of these cases, there is reduced understanding of the 'how' of system behaviour. However, here we need to balance prospective responsibility (i.e. the obligations we have as a result of our role) against retrospective responsibility (i.e. accountability for outcomes). For as long as the behaviour is, in principle, understandable, then it is reasonable to maintain that it is incumbent upon the person whose role to is to analyse the system to understand that system; and similarly with the analysis of credible hazards.

5 Responsibility Gaps and Autonomous Systems

There are no *established* safety engineering frameworks for autonomous systems. The test coverage measures discussed above do not yet exist for autonomous systems.

By their very nature, autonomous systems are developed to operate in complex environments. Here, it is not possible to pre-define what would be correct and safe system behaviour, and to pre-program accordingly. The machine itself must process, interpret, and action large, dynamic data sets. With the relationship (or transfer function) between the inputs and outputs for such systems being difficult, if not impossible, to describe algorithmically, machine learning is deployed to 'teach' the system its transfer function. For example, in a personalised healthcare solution, the clinical advice of an autonomous system depends on conditions, behaviours, constraints, and preferences that are learnt by the machine at runtime that cannot be predicted prior to deployment.

Causal control, over the 'what' of system behaviour, becomes more challenging. Though engineers and designers still determine what the system's capabilities are and what it can be used for, they are not always in control of how these might change as a result of the machine's own learning. The epistemic dimension - moral responsibility for the 'how' of system behaviour - is even more problematic. With autonomy, it becomes exceedingly difficult to understand how a system gets from input to output. This presents a difficulty for the existing safety assurance paradigm, given that a large segment of the safety argument is effectively the confidence in the 'how'.

One problem is the distinction between system-type and individual examples of the system. In philosophy, this would be known as a type-token distinction. An autonomous system, or system-type, will have a 'correct' design (i.e. the AI engine), but each system, or system-token, will differ in its actual behaviour

once it starts learning 'in the wild' (i.e. based on real-world data). Not only is it difficult to foresee what learning errors might occur, it will also be difficult to distinguish a design error from a learning error. All of this serves to undermine the extent to which a safety engineer can be responsible for - and mitigate against - subsequent hazardous behaviour.

We should also consider the causal influence of the users (and society) from whom the system has learnt the unsafe behaviours. Here, in conditions of uncertainty, it would seem that there is some duty, a prospective moral responsibility, incumbent upon all parties - engineers and users - to ensure that the system is provided with exemplary training data. If this still leads to inexplicable negative outcomes, then an accountability gap remains.

Two further problems for moral responsibility bear consideration, both of which complicate the line between causal control (over the 'what') and epistemic control (over the 'how'). The first is emergent behaviour. A learning system might start to generate new behaviours. Here, the design (the 'how') equips the system with the ability to change what it can do and be used for (the 'what'). Changes in the system's behaviour occur because of what the system has learnt by way of its learning algorithms, as opposed to change that is directly influenced by the system's design.

This links to the second problem: trade-offs. An autonomous system might learn original and highly effective skills, but if we have a limited scope of understanding as to what 'safe' looks like, we might unnecessarily or unwittingly constrain this behaviour. With the personalised healthcare system, for example, safeguards can be put in place to mitigate some possible consequences, such as recommendations that physical exercises do not exceed certain thresholds. But these safeguards might reduce the machine's ability to perform tasks that lead to even safer or more effective actions that the designers did not, or could not, foresee. Given the peculiarities of the situation, non-compliance or improvisation might lead to safer outcomes than following established safety procedures. This raises the question of the extent to which designers can deliberate about the possible consequences of the actions of an autonomous system in emergency and novel situations.

6 Conclusions and Future Considerations

One thing is clear from the foregoing discussion: the safety engineering community is in urgent need of robust and open deliberation about what is deemed as sufficient control over the 'what' and the 'how' of autonomous system behaviour.

Substantive issues that we have raised, and which we think bear consideration with respect to the moral responsibilities on the engineering community (as well as on users and society) are as follows. If we can only assure system-types, how should we navigate the unpredictability of system-tokens? To what level of confidence should we be able to distinguish a learning error from a design error? Is it possible to distinguish between a learning error and an original solution to a problem? How should we account for emergent behaviour? What standards

should we use to reconcile trade-offs, for example between explainability and effectiveness?

There is also a need to focus on cultural acceptance: an obligation on autonomous systems engineers to show these systems' safety advantage over human-controlled systems. Part of this includes moves to increase the explainability of such systems, such that the system can explain the decision made, or at least why its course of action led to a 'more safe' outcome than any of the alternatives.

These considerations locate the key areas of uncertainty about, and diminished control over, the behaviour of autonomous systems. We believe that these considerations should feed into discussions about the development of such systems from moral and liability perspectives.

Such discussion is both a question of retrospective moral responsibility: how far system failure can be traced back to design, engineering, and user fault. But it is also demand upon prospective moral responsibility: how to determine confidence in autonomous systems in order to assure their future safety.

References

1. Lucas, J.R.: Responsibility. (1993)
2. Talbert, M.: Moral Responsibility. Polity Press (2015)
3. Aristotle: The Nichomachean Ethics. Oxford University Press (2009)
4. Strawson, P.F.: Freedom and resentment. In Watson, G., ed.: Proceedings of the British Academy, Volume 48: 1962. Oxford University Press (1962) 1–25
5. Eshleman, A.: Moral responsibility. (2016) Accessed: 20 June 2018.
6. Fischer, J.M., Ravizza, M.: Responsibility and Control: A Theory of Moral Responsibility. Cambridge University Press (2000)
7. Hyman, J.: Action, Knowledge, and Will. Oxford University Press (2015)
8. Brink, D.O., Nelkin, D.: Fairness and the architecture of responsibility. Oxford Studies in Agency and Responsibility **1** (2013) 284–313
9. Matthias, A.: The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology **6**(3) (2004) 175–183
10. Noorman, M.: Computing and moral responsibility. (2018) Accessed: 08 Mar 2018.
11. Jonas, H.: The Imperative of Responsibility: In Search of an Ethics for the Technological Age. University of Chicago Press (1985)